3
Publishing Technique

As discussed in Chapter 2, annotations can be extracted from audio, text, and visual features. The extraction of text features from the audio layer is the approach generally used when the multimedia asset is rich in speech, achieving good indexing and retrieval results for multimedia content. For this reason, at the end of Chapter 2, related works in this direction were surveyed.

Unfortunately, the indexing and retrieval systems presented in Chapter 2.5 were built to be consumed via a Web form, which generates a barrier to search engines that prevents indexing multimedia content. Therefore, we realized that a suitable format to publish this type of textual annotations is needed.

With our approach, we aim at transforming the multimedia content so that it becomes visible to search engines. For this reason, we must make this content available in the most suitable format to be understood and processed by both the search engines and human beings, without duplicating the content.

In this chapter, we describe the publishing process we developed. We explain in detail the importance of each step of the process and its impact on the indexing and retrieval of multimedia assets.

3.1
Overview of the Publishing Technique

The publishing technique aims at publishing multimedia assets – audio and video – on the Web in a way that improves the ability of search engines to index and retrieve the multimedia assets. The technique includes the use of Semantic Web technology that recent search engines can explore to better index and retrieve the multimedia assets. It also guides users through the publication process.

Today, most multimedia content indexing and retrieval systems, such as [3, 26, 27, 29], creates indexes for both metadata and the transcription. To consume this information, a Web interface is created for searching and browsing the multimedia assets, with little concern to publishing this content in a suitable format to be consumed by the standard search engines. Consequently, the multimedia assets remain invisible (in the Deep Web) and cannot be consumed by users across the Web. We believe that the creation of static Web pages [35] with Semantic Web Standards [33] is the way to make multimedia content visible to the search engines. Figure 8 presents the publishing technique architecture.

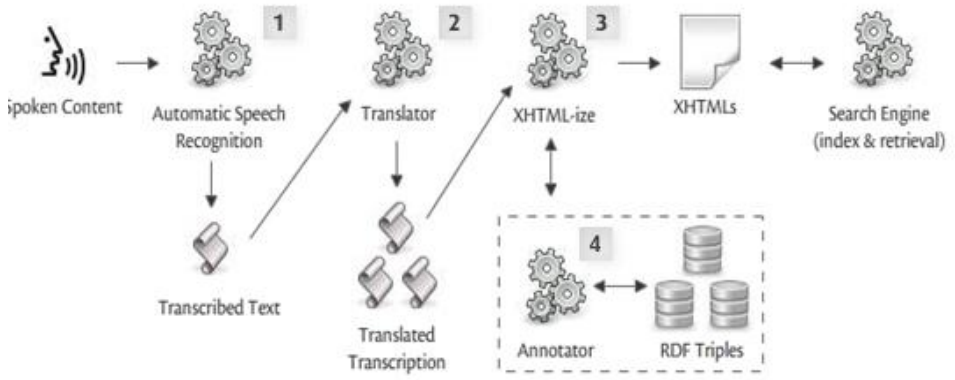

Figure 8 Publishing technique architecture.

The proposed process consists of the following stages: automatic speech recognition, translation, XHTML-ize and annotation. The process starts by exploring a straightforward strategy: since multimedia assets are opaque to standard search engines, the process proposes to use the associated text to help standard search engines index and retrieve the multimedia assets. Stage 1 extracts a transcription from a multimedia asset by processing the associated closed caption file, if any, or by a speech recognition tool (this is not the focus of the dissertation, though). Stage 2 translates the transcription to other languages, which allows non-native speakers of the original language of the multimedia assets to understand the content. For each multimedia asset, stage 3 creates an XHTML page that allows search engines to find and retrieve segments of the multimedia asset that correspond to paragraphs of the associated XHTML page. Stage 4 includes an annotation step that recognizes entities in the text associated with an A/V asset and includes references to entries in DBPedia, Freebase,

Schema.org or any other Tripleset available through a SPARQL endpoint. The annotations are included in the text as RDFa markup.

In the following sections will be detailed these steps.

3.2 Stage 1 - Automatic speech recognition

This stage consists of extracting the transcription from a multimedia asset by processing the associated closed caption file, if any, or it is suggested that an automatic speech recognition service (ASR) be used for transcribing the spoken content [30, 31]. An ASR takes as an input a multimedia asset, rich in speech, and generates a timed-aligned transcription.

The goal of performing this process is to extract some semantic meaning of the spoken content. As mentioned in [12], the transcription is one form of extracting semantics from a multimedia asset, which can be presented in multiple forms, e.g., data from sound tracks, text extracted from the image frames, image frames and spoken words in the audio layer.

Since we want to take advantage of the extensive research involving information retrieval, we transform the spoken search problem into a text search problem. Figure 10 shows an example of a script.

3.3 Stage 2 - Translation

During the translation stage, the publishing technique transforms the transcription to other languages, if desired, in order to reach other user populations.

Indeed, once the spoken content is in a text form, text-based search engines are able to index the content. Nevertheless, many Web users remain unable to consume the content since the content is not available in their native language. The main reason is that some Web users can read documents written in different languages, but they cannot formulate an ideal keyword search query because they cannot formulate search terms comparable to those of a native speaker [32].

The “multilinguality” of Web content provides opportunities for users to directly access and use previously incomprehensible sources of Web information. However, Web users find it difficult to take advantage of these opportunities when the online information access systems are monolingual [37].

It is well known that the Web is composed mostly of content written in English — about 56.4%. Furthermore, most of the research on information retrieval focused on documents written in English. However, today, production of content in other languages is growing. For example, we see that Web content in German in 2002 was approximately 7.7%, in French 5.6%, in Spanish 3% and in Portuguese 1.5%. This attracted more and more of the attention of the academic community focused on information retrieval [5].

Cross-lingual information retrieval (CLIR) arises as a solution to meet the needs of dealing with content in different languages. The aim of CLIR is to allow user to make queries in one language and retrieve documents in one or more other languages [38]. Then, the retrieved pages are translated into the language used for the query. For example, a user can make a query in Portuguese about “receitas de frango” and receive documents in Catalan about “pollastre ams verduretes”, which is “chiken recipe”.

CLIR techniques suffer from some drawbacks because each word in the search is translated using a bilingual dictionary. It was found in the literature that using a bilingual dictionary results in an inaccurate query. The loss of semantics in the query is mainly because each query term has several meanings in the dictionary. In other words, each term have one or more possible translations in the dictionary. As a result, we have a translated query, but not with the same semantics as the original, which may lead to the retrieval of documents unrelated to the initial search.

Machine translation (MT) is one of the components of a CLIR system. MT aims at automating the translation process. This technology includes analyzing and understanding information in one language to expressing it in another language [38, 37]. In [36], the authors present some of the possible advantages

[5] http://www.netz-tipp.de/languages.html

of document translation. The most important is that, by translating the document, there are more chances for translating a word correctly, that is, there are more text in the document than a query. This helps improving the translation using the relations between the words in the full document.

Although considerable research has been carried out in these areas, there is still a lack of services that provide CLIR and MT. As a solution, in 2008, Google was the first and the only search engine to offer the service of cross-language search. This service was the first implementation of the theory of cross-language information retrieval (CLIR) to be put in practice.

Our technique uses a machine translation (MT) from the Google language tools. The Google Translator API (http://code.google.com/apis/language/) is used to translate the transcription into several languages. The MT service of Google allows instant translation between 58 different languages. In Chapter 5, we discuss how this step impacts multimedia content retrieval. The goal of using the Google translate service is to make the information universally accessible and useful, regardless of the language in which it was written.

It is good to mention that a MT is error-prone. Thus, it would be interesting as a future step to perform a manually revision to correct inappropriate translation in translated document, that would help to increase the precision of MT became almost equivalent to that of human translation [42].

3.4 Stage 3 - XHTML-ize

The XHTML-ize stage transforms a plain text script into a XHTML (eXtensible Hypertext Markup Language) page. This stage generates static Web pages for each asset and converts each time-aligned excerpt of the transcription from a spoken content into sections (“div elements”), where each section contains a hyperlink (the “a element”) that points to the exact segment of the spoken content where the speech occurs (see Figure 18).

The title and description of the spoken content is also added to the XHTML page; this information is configured inside a div section that contains a heading

tag (*h1 element*) for title and a paragraph tag (*p element*) for the description. Figure 17 illustrates the process.

It is important to note that the construction of the XHTML pages follows the recommendation of the World Wide Web consortium W3C, established in the Web Content Accessibility Guidelines (WCAG) [41], which helps making the Web content accessible to people with impairments.

The task of searching information in hours of audio/video content is time-consuming. Therefore, linking text segments with video segments helps locating the relevant information or topic of one's interest covered in the asset.

We can see the utility of this strategy in the following scenario. Suppose that a user is using a mobile device, such as smartphone, PDA, Ultra-Mobile PC or Tablet PC. The goal is to find some relevant information within an hour of a spoken content. The first approach would be to view all the asset until finding the desire content, but this can take a long time. This approach has several drawbacks: downloading the complete spoken content over the Internet, or perhaps over a wireless network to a mobile device, would require the use of much more bandwidth and will generate a high cost to the user and congestion in the network. Other problems that may arise are inefficient use of the limited resources of the mobile device and the loss of the user's attention.

The strategy we implemented - linking text segments with video segments - optimizes the search of the desired point within the asset, maximizing resource usage and retaining user's attention.

3.5 Stage 4 - Annotation

This stage consists in the recognition and annotation of entities in the transcription. It takes as input the Web document generated in the previous stage, which contains the transcription of the spoken content, and outputs an enriched Web document.

In the past, semantic (machine-readable) data was distributed separately, using Resource Description Framework (RDF). Resource Description

Framework-in-attributes (RDFa) was created as a solution to distribute machine-readable data with embedded semantics by providing a set of attributes specified for XHTML 1.1. In this sense, RDFa helps distributing a single version of the document for both machine and human consumption [33]. A recent survey, released by Yahoo!, showed that RDFa has had a growth 510% in 2010 [39].

Enriching Web documents with RDFa allows search engines to better understand the content on the Web pages. The engines can then provide additional information to help Web users identify relevant information [40]. One application of RDFa is Google's Rich Snippets[6], which help users find Web pages by showing relevant information about what the user will find when he accesses the Web page. Google Rich Snippets are the result of applying Google algorithms that highlight structured data embedded in Web Pages.

To achieve the goals of this stage, we first recognize entities in the Web Document. We use entity detection and name resolution, which are subtasks of the information extraction area (IE), whose target is to extract structured information from unstructured or semi-structured machine-readable documents. Entity detection tries to recognize known entities, such as temporal expressions, places, people and organization names. Name resolution tries to find when two recognized entities refer to the same real-world entity. In our approach, we use the Spotlight[7] Web Service to detect entities in the Web document and include references to DBPedia[8], Freebase[9], Schema.org[10] or any other Tripleset available through a SPARQL endpoint resource. That is, we are linking unstructured information sources to the Linked Open Data cloud (LOD) through DBpedia.

Finally, we include annotations in the Web document as RDFa markup. As mentioned before, such semantic annotations are used by recent search engines to improve page ranking. Furthermore, they provide the final user with additional information about the concepts addressed in the transcription asset.

[6] http://www.google.com/webmasters/tools/richsnippets

[7] http://dbpedia.org/spotlight

[8] http://dbpedia.org/About

[9] http://www.freebase.com/

[10] http://schema.org/